Bayesian Analysis (2020) **15**, Number 3, pp. 737–758

Bayesian Bandwidth Test and Selection for High-dimensional Banded Precision Matrices

Kyoungjae Lee* and Lizhen Lin†

Abstract. Assuming a banded structure is one of the common practice in the estimation of high-dimensional precision matrices. In this case, estimating the bandwidth of the precision matrix is a crucial initial step for subsequent analysis. Although there exist some consistent frequentist tests for the bandwidth parameter, bandwidth selection consistency for precision matrices has not been established in a Bayesian framework. In this paper, we propose a prior distribution tailored to the bandwidth estimation of high-dimensional precision matrices. The banded structure is imposed via the Cholesky factor from the modified Cholesky decomposition. We establish strong model selection consistency for the bandwidth as well as consistency of the Bayes factor. The convergence rates for Bayes factors under both the null and alternative hypotheses are derived which yield similar order of rates. As a by-product, we also propose an estimation procedure for the Cholesky factors yielding an almost optimal order of convergence rates. Two-sample bandwidth test is also considered, and it turns out that our method is able to consistently detect the equality of bandwidths between two precision matrices. The simulation study confirms that our method in general outperforms the existing frequentist and Bayesian methods.

MSC 2010 subject classifications: Primary 62H15, 62C10; secondary 62F05.

Keywords: precision matrix, bandwidth selection, Cholesky factor, convergence rates of Bayes factor.

1 Introduction

Estimating a large covariance or precision matrix is a challenging task in both frequentist and Bayesian frameworks. When the number of variables p is larger than the sample size n, the traditional sample covariance matrix does not provide a consistent estimate of the true covariance matrix (Johnstone and Lu, 2009), and the inverse Wishart prior leads to the posterior inconsistency (Lee and Lee, 2018b). To overcome this issue, various restricted classes of matrices have been investigated such as the bandable matrices (Bickel and Levina, 2008; Cai et al., 2010; Hu and Negahban, 2017; Banerjee and Ghosal, 2014; Lee and Lee, 2018a), sparse matrices (Cai and Zhou, 2012a; Banerjee and Ghosal, 2015; Xiang et al., 2015; Cao et al., 2019) and low-dimensional structural matrices (Fan et al., 2008; Cai et al., 2015; Pati et al., 2014; Gao and Zhou, 2015). In this paper, we focus on *banded precision matrices*, where the banded structure is encoded via the Cholesky factor of the precision matrix. We are in particular interested in the estimation

*Department of Statistics, Inha University, South Korea, leekjstat@gmail.com

†Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, USA, lizhen.lin@nd.edu

 https://doi.org/10.1214/19-BA1167

of the bandwidth parameter and construction of Bayesian bandwidth tests for one or two banded precision matrices. Inference of the bandwidth is of great importance for detecting the dependence structure of ordered data. Moreover, it is a crucial initial step for subsequent analysis such as linear or quadratic discriminant analysis.

Bandwidth selection of the high-dimensional precision matrices has received increasing attention in recent years. An et al. (2014) proposed a test for bandwidth selection, which is asymptotically normal under the null hypothesis and has a power tending to one. Based on the proposed test statistics, they constructed a backward procedure to detect the true bandwidth by controlling the familywise errors. Cheng et al. (2017) suggested a bandwidth test without assuming any specific parametric distribution for the data and obtained a result similar to that of An et al. (2014).

In the Bayesian literature, Banerjee and Ghosal (2014) studied the estimation of bandable precision matrices which include the banded precision matrix as a special case. They derived the posterior convergence rate of the precision matrix under the G-Wishart prior (Roverato, 2000). Lee and Lee (2018a) considered a similar class to that of Banerjee and Ghosal (2014), but assumed bandable Cholesky factors instead of bandable precision matrices. They showed the posterior convergence rates of the precision matrix as well as the minimax lower bounds. In both works, posterior convergence rates were obtained for a given (fixed) bandwidth, and the posterior mode was suggested as a bandwidth estimator in practice. However, no theoretical guarantee is provided for such estimators. Further, no Bayesian bandwidth test exists for one- or two-sample problems.

This gap in the literature motivates us to investigate theoretical properties related to the general problem of bandwidth test and selection, and propose estimators or tests with theoretical guarantees. In this paper, we use the modified Cholesky decomposition of the precision matrix and assume banded Cholesky factors. The induced precision matrix also has banded structure. The key difference from Lee and Lee (2018a) is on the choice of prior distributions which will be introduced in Section 2.3. In addition, we focus on bandwidth selection and tests, while Lee and Lee (2018a) mainly studied the convergence rates of the precision matrix for a given or fixed bandwidth.

There are two main contributions of this paper. First, we suggest a Bayesian procedure for banded precision matrices and prove the bandwidth selection consistency (Theorem 3.1) and consistency of the Bayes factor (Theorem 3.2). To the best of our knowledge, our work is the first that has established the bandwidth selection consistency for precision matrices under a Bayesian framework, which implies that the marginal posterior probability for the true bandwidth tends to one as $n \to \infty$. Cao et al. (2019) proved strong model selection consistency for the sparse directed acyclic graph (DAG) models, but their method is not applicable to the bandwidth selection problem since it is not adaptive to the unknown sparsity. Lee et al. (2018) also proved the strong model selection consistency for the sparse DAG models, but they used the fractional likelihood approach, which cannot be applicable to the Bayesian testing. Second, we also prove the consistency of the Bayes factor for two-sample bandwidth testing problem (Theorem 3.3) and derived the convergence rates of the Bayes factor under both the null and alternative hypotheses. Our method is able to consistently detect the equality of bandwidths between two different precision matrices. To the best of our knowledge,

this is also the first consistent two-sample bandwidth test result in both frequentist and Bayesian literature. The existing literature (frequentist) focused only on the one-sample bandwidth testing (An et al., 2014; Cheng et al., 2017).

The rest of the paper is organized as follows. Section 2 introduces the notations, model, priors and assumptions used. Section 3 describes main results of this paper: bandwidth selection consistency and convergence rates of one- and two-sample bandwidth tests. Simulation study and real data analysis are presented in Section 4 to show the practical performance of the proposed method. In Section 5, concluding remarks and topics for the future work are given. The supplementary material (Lee and Lin, 2019) includes a result on the nearly optimal estimation of the Cholesky factors, and proofs of main results.

2 Preliminaries

2.1 Notations

For any real numbers a and b, we denote $a \wedge b$ and $a \vee b$ as the minimum and maximum of a and b, respectively. For any sequences a_n and b_n, we denote $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. We write $a_n \lesssim b_n$, or $a_n = O(b_n)$, if there exists an universal constant $C > 0$ such that $a_n \leq C b_n$ for any n. We define vector ℓ_2- and ℓ_∞-norms as $\|a\|_2 = (\sum_{j=1}^p a_j^2)^{1/2}$ and $\|a\|_\infty = \max_{1 \leq j \leq p} |a_j|$ for any $a = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$. For a matrix A, the matrix ℓ_∞-norm is defined as $\|A\|_\infty = \sup_{\|x\|_\infty = 1} \|Ax\|_\infty$. We denote $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ as the minimum and maximum eigenvalues of A, respectively.

2.2 Gaussian Models

We consider a Gaussian model

$$X_1, \ldots, X_n \mid \Omega_n \overset{i.i.d.}{\sim} N_p(0, \Omega_n^{-1}), \tag{1}$$

where $\Omega_n = \Sigma_n^{-1}$ is a $p \times p$ precision matrix and $X_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$ for all $i = 1, \ldots, n$. For any positive definite matrix Ω_n, there exist unique lower triangular matrix $A_n = (a_{jl})$ and diagonal matrix $D_n = diag(d_j)$ such that

$$\Omega_n = (I_p - A_n)^T D_n^{-1} (I_p - A_n), \tag{2}$$

where $a_{jj} = 0$ and $d_j > 0$ for all $j = 1, \ldots, p$, by the modified Cholesky decomposition (MCD). We call A_n the *Cholesky factor*. Define k as the *bandwidth* of a matrix if the off-diagonal elements of the matrix farther than k from the diagonal are all zero. If the bandwidth of the Cholesky factor is k, model (1) can be represented as

$$\begin{aligned} X_{i1} \mid d_1 &\overset{i.i.d.}{\sim} N(0, d_1), \\ X_{ij} \mid a_j^{(k)}, d_j, k &\overset{ind}{\sim} N\Big(\sum_{l=(j-k)_1}^{j-1} X_{il} a_{jl}, d_j \Big), \quad j = 2, \ldots, p \end{aligned} \tag{3}$$

for all $i = 1, \ldots, n$, where $a_j^{(k)} = (a_{jl})_{(j-k)_1 \le l \le j-1} \in \mathbb{R}^{k_j}$, $(j-k)_1 = 1 \vee (j-k)$ and $k_j = k \wedge (j-1)$. The above representation enables us to adopt priors and techniques in the linear regression literature.

We are interested in the consistent estimation and hypothesis test of the bandwidth k of the precision matrix. From the decomposition (2), *the bandwidth of A_n is k if and only if the bandwidth of Ω_n is k.* Thus, we can infer the bandwidth of the precision matrix by inferring that of the Cholesky factor.

2.3 Prior Distribution

Let $\tilde{X}_j \in \mathbb{R}^n$ and $\mathbf{X}_{j(k)} \in \mathbb{R}^{n \times k_j}$ be sub-matrices consisting of jth and $(j-k)_1, \ldots, (j-1)$th columns of $\mathbf{X}_n = (X_1^T, \ldots, X_n^T)^T \in \mathbb{R}^{n \times p}$, respectively. We suggest the following prior distribution

$$a_j^{(k)} \mid d_j, k \overset{ind}{\sim} N_{k_j}\left(\widehat{a}_j^{(k)}, \frac{d_j}{\gamma}\left({\mathbf{X}_{j(k)}}^T \mathbf{X}_{j(k)}\right)^{-1}\right), \quad j = 2, \ldots, p, \tag{4}$$

$$\pi(d_j) \overset{i.i.d.}{\propto} d_j^{\tau n/2 - 1}, \quad j = 1, \ldots, p, \tag{5}$$

$$k \sim \pi(k), \quad k = 0, 1, \ldots, R_n \tag{6}$$

for some positive constants γ, τ and positive sequence R_n, where $\widehat{a}_j^{(k)} = ({\mathbf{X}_{j(k)}}^T \times \mathbf{X}_{j(k)})^{-1} {\mathbf{X}_{j(k)}}^T \tilde{X}_j$. The conditional prior distribution for $a_j^{(k)}$ is a version of the Zellner's g-prior (Zellner, 1986; Martin et al., 2017) in the linear regression literature. Note that model (3) is equivalent to $\tilde{X}_j \mid a_j^{(k)}, d_j, k \sim N_n\left(\mathbf{X}_{j(k)} a_j^{(k)}, d_j I_n\right)$. Due to the conjugacy, it enables us to calculate the posterior distribution in a closed form up to some normalizing constant. The prior for d_j is carefully chosen to reduce the posterior mass towards large bandwidth k. We emphasize here that one can use the usual non-informative prior $\pi(d_j) \propto d_j^{-1}$, but necessary conditions for the main results in Section 3 should be changed. This issue will be discussed in more details in the next paragraph. We assume the prior $\pi(k)$ to have the support on $0, 1, \ldots, R_n$. We will introduce condition (A4) for $\pi(k)$ and the hyperparameters in Section 2.4, and show that $\pi(k) \propto 1$ is enough to establish the main results in Section 3.

The priors (4)–(6) lead to the following joint posterior distribution,

$$\begin{aligned}
a_j^{(k)} \mid d_j, k, \mathbf{X}_n &\overset{ind}{\sim} N_{k_j}\left(\widehat{a}_j^{(k)}, \frac{d_j}{1+\gamma}\left({\mathbf{X}_{j(k)}}^T \mathbf{X}_{j(k)}\right)^{-1}\right), \quad j = 2, \ldots, p, \\
d_j \mid k, \mathbf{X}_n &\overset{ind}{\sim} IG\left(\frac{(1-\tau)n}{2}, \frac{n}{2}\widehat{d}_j^{(k)}\right), \quad j = 1, \ldots, p, \\
\pi(k \mid \mathbf{X}_n) &\propto \pi(k) \prod_{j=2}^{p} \left(1 + \frac{1}{\gamma}\right)^{-\frac{k_j}{2}} (\widehat{d}_j^{(k)})^{-\frac{(1-\tau)n}{2}}, \quad k = 0, 1, \ldots, R_n,
\end{aligned} \tag{7}$$

provided that $\tau < 1$, where $\widehat{d}_j^{(k)} = \tilde{X}_j^T (I - \tilde{P}_{jk}) \tilde{X}_j / n$ and $\tilde{P}_{jk} = \mathbf{X}_{j(k)} ({\mathbf{X}_{j(k)}}^T \mathbf{X}_{j(k)})^{-1} \times {\mathbf{X}_{j(k)}}^T$. The marginal posterior $\pi(k \mid \mathbf{X}_n)$ consists of two parts: the penalty on the model

size, $\pi(k)\prod_{j=2}^{p}(1+1/\gamma)^{-k_j/2}$, and the estimated residual variances, $\prod_{j=2}^{p}(\widehat{d}_j^{(k)})^{-(1-\tau)n/2}$. Thus, priors (4) and (5) naturally impose the penalty term $\prod_{j=2}^{p}(1+1/\gamma)^{-k_j/2}$ for the marginal posterior $\pi(k \mid \mathbf{X}_n)$.

The effect of prior $\pi(d_j) \propto d_j^{\tau n/2-1}$ appears in marginal posterior for k. Compared with the prior $\pi(d_j) \propto d_j^{-1}$, it produces the term $(\widehat{d}_j^{(k)})^{-(1-\tau)n/2}$ instead of $(\widehat{d}_j^{(k)})^{-n/2}$. It reduces the posterior ratio $\pi(k+1 \mid \mathbf{X}_n)/\pi(k \mid \mathbf{X}_n)$ since $\widehat{d}_j^{(k)}$ decreases as k grows, so it reduces the relative posterior mass towards large bandwidth k. We conjecture that, at least for our prior choice of $\pi(a_j^{(k)} \mid d_j, k)$ with a constant $\gamma > 0$, this power adjustment of $\widehat{d}_j^{(k)}$ is essential to prove the selection consistency for k. Suppose we use the prior $\pi(d_j) \propto d_j^{-1}$. Similar to the proof of Theorem 3.1, to obtain the selection consistency, we will use the inequality

$$\pi(k \mid \mathbf{X}_n) \quad \leq \quad \frac{\pi(k \mid \mathbf{X}_n)}{\pi(k_0 \mid \mathbf{X}_n)} \;=\; \frac{\pi(k)}{\pi(k_0)} \prod_{j=2}^{p} \left(1+\frac{1}{\gamma}\right)^{-\frac{k_j-k_{0j}}{2}} \left(\frac{\widehat{d}_j^{(k)}}{\widehat{d}_j^{(k_0)}}\right)^{-\frac{n}{2}}, \tag{8}$$

and show that the expectation of the right hand side term converges to zero for any $k \neq k_0$ as $n \to \infty$, where k_0 is the true bandwidth. Note that unless $\pi(k_0 \mid \mathbf{X}_n)$ shrinks to zero, the inequality causes only a constant multiplication. The most important task is dealing with the last term in (8), $(\widehat{d}_j^{(k)}/\widehat{d}_j^{(k_0)})^{-n/2}$. Concentration inequalities for chi-square random variables (for examples, see Lemma 3 in Yang et al. (2016) and Lemma 4 in Shin et al. (2018)) suggest an upper bound $p^{\alpha(k_j-k_{0j})}$ with high probability for any $2 \leq j \leq p$, $k > k_0$ and some constant $\alpha > 0$. In this case, the hyperparameter γ should be of order $p^{-\alpha'}$ for some constant $\alpha' > 2\alpha$ to make the right hand side in (8) converge to zero. Then, with the choice $\gamma \asymp p^{-\alpha'}$, condition (A2), which will be introduced in Section 2.4, should be modified by replacing $1/n$ with $(\log p)/n$ to achieve the selection consistency. In summary, the main results in this paper still hold for the prior $\pi(d_j) \propto d_j^{-1}$, but it requires stronger conditions due to technical reasons. We state the results using prior (5) to emphasize that the bandwidth selection problem essentially requires weaker condition than the usual model selection problem.

Remark. If we adopt the fractional likelihood (Martin et al., 2017), we can achieve the selection consistency (Theorem 3.1) with the prior $\pi(d_j) \propto d_j^{-1}$ instead of (5) under similar conditions in Theorem 3.1. However, with the fractional likelihood, we cannot calculate the Bayes factor which is crucial to describe the Bayesian test results in Sections 3.2 and 3.3.

Remark. There are two consequences by using the data-dependent mean $\widehat{a}_j^{(k)}$. First, we can avoid assuming an upper bound condition for $\|\mathbf{X}_{j(k_0)}a_{0,j}^{(k_0)}\|_2$ or $\|a_{0,j}^{(k_0)}\|_2$, where $a_{0,j}^{(k_0)} = (a_{0,jl})_{(j-k_0)_1 \leq l \leq j-1}$ denotes the sub-vector of the true Cholesky factor. An upper bound condition is required if we adopt the Zellner's g-prior with zero mean (Shang and Clayton, 2011), e.g., Yang et al. (2016) assumed $\|\mathbf{X}_{j(k_0)}a_{0,j}^{(k_0)}\|_2^2 \leq \gamma^{-1}d_{0j}\log p$ in order to prove selection consistency for the regression coefficient vector. Second, we do not need to assume the so-called information paradox of Zellner's g-prior (Liang et al., 2008), which corresponds to $\gamma = p^{-2c}$ for some $c \geq 1/2$ in our notation. In this paper, we assume γ is a constant satisfying some conditions in Section 2.4.

2.4 Assumptions

We denote Ω_{0n} as the true precision matrix whose MCD is given by $\Omega_{0n} = (I_p - A_{0n})^T D_{0n}^{-1}(I_p - A_{0n})$. Let $\mathbb{P}_0$ and $\mathbb{E}_0$ be the probability measure and expectation corresponding to model (1) with Ω_{0n}. For the true Cholesky factor $A_{0n} = (a_{0,jl})$, we denote k_0 as the true bandwidth. We introduce conditions (A1)–(A4) for the true precision matrix and priors (4)–(6):

(A1). Assume that p increases to the infinity as $n \to \infty$. Furthermore, there exist positive sequences $\epsilon_{0n} \leq 1$ and $\zeta_{0n} \geq 1$ such that $\epsilon_{0n} \leq \lambda_{\min}(\Omega_{0n}) \leq \lambda_{\max}(\Omega_{0n}) \leq \zeta_{0n}$ for every $n \geq 1$ and $\zeta_{0n} \log p/\epsilon_{0n} = o(n)$.

(A2). For a given positive constant $\tau \in (0, 0.4]$ in prior (5), there exists a positive constant M_{bm} such that for every $n \geq 1$,

$$\min_{j,l:a_{0,jl}\neq 0} |a_{0,jl}|^2 \quad \geq \quad \frac{10M_{\text{bm}}}{\tau(1-\tau)\,n}\,\frac{\zeta_{0n}}{\epsilon_{0n}}.$$

(A3). The sequence R_n in prior (6) satisfies $k_0 \leq R_n \leq \min\left\{n\tau\epsilon_1(1+\epsilon_1)^{-1}, (1-\epsilon_2)p\right\}$ for some small $0 < \epsilon_1, \epsilon_2 < 1$ and all sufficiently large n.

(A4). For given positive constants γ and $\tau \in (0, 0.4]$ in priors (4) and (5), assume that

$$\sum_{k>k_0} \frac{\pi(k)}{\pi(k_0)} \{C_{\gamma,\tau}\}^{-(k-k_0)\cdot(p-\frac{k+k_0+1}{2})} \quad = \quad o(1), \tag{9}$$

$$\sum_{k<k_0} \frac{\pi(k)}{\pi(k_0)} \{C_{\gamma,M_{\text{bm}}}\}^{(k_0-k)\cdot(p-\frac{k+k_0+1}{2})} \quad = \quad o(1), \tag{10}$$

where $C_{\gamma,\tau} = \left\{(1+\gamma^{-1})\cdot\tau(1+\epsilon_1)^{-1}\right\}^{1/2}$ and $C_{\gamma,M_{\text{bm}}} = 2\,(1+\gamma^{-1})^{1/2}\exp(-M_{\text{bm}})$.

Now, let us describe the above conditions in more detail. The bounded eigenvalue condition for the true precision matrix is common in the high-dimensional precision matrix literature (Banerjee and Ghosal, 2014, 2015; Xiang et al., 2015; Ren et al., 2015; Lee et al., 2018). We allow that $\epsilon_{0n} \to 0$ and $\zeta_{0n} \to \infty$ as $n \to \infty$, so condition (A1) is much weaker than the condition in the above literature, which assumes $\epsilon_{0n} = \zeta_{0n}^{-1} = \epsilon_0$ for some small constant $\epsilon_0 > 0$. Cao et al. (2019) also allowed diverging bounds, but assumed that $\zeta_{0n} = \epsilon_{0n}^{-1}$ and $(\log p/n)^{1/2-1/(2+t)} = o(\epsilon_{0n}^4)$ for some $t > 0$. If we assume that $\zeta_{0n} = \epsilon_{0n}^{-1}$, then condition (A1) implies that $\log p/n = o(\epsilon_{0n}^2)$, which is much weaker than the condition used in Cao et al. (2019).

Condition (A2) is called the *beta-min condition*. If we assume that $\epsilon_{0n} = O(1)$ and $\zeta_{0n} = O(1)$, in our model it only requires the lower bound of the nonzero elements to be of order $O(1/\sqrt{n})$. In the sparse regression coefficient literature, the lower bound of the nonzero coefficients is usually assumed to be $\sqrt{\log p/n}$ up to some constant (Castillo et al., 2015; Yang et al., 2016; Martin et al., 2017; Lee et al., 2018). Here, the $\sqrt{\log p}$ term can be interpreted as a price coming from the absence of information on the zero-pattern. Condition (A2) reveals the fact that, under the banded assumption, we do not need to pay this price anymore.

Condition (A3) ensures that the true bandwidth k_0 lies in the support of $\pi(k)$. Note that $k_0 \leq (1-\epsilon_2)p$ is not an additional condition because the support of bandwidth should be smaller than p. The condition $k_0 \leq n\tau\epsilon_1(1+\epsilon_1)^{-1}$ is needed for the selection consistency, which holds if we choose $R_n = C \vee \{n\tau\epsilon_1(1+\epsilon_1)^{-1}\}$ for some large constant $C > 0$. Although this is slightly stronger than the condition $k_0 \leq n-4$ in An et al. (2014), it is much weaker than those in other works. For examples, Banerjee and Ghosal (2014) assumed $k_0^5 = o(n/\log p)$ for the consistent estimation of precision matrix, and Cheng et al. (2017) assumed $k_0 = O([n/\log p]^{1/2})$ for theoretical properties.

The equations (9) and (10) in condition (A4) guarantee $\mathbb{E}_0[\pi(k > k_0 \mid \mathbf{X}_n)] = o(1)$ and $\mathbb{E}_0[\pi(k < k_0 \mid \mathbf{X}_n)] = o(1)$, respectively. Here we give some examples for $\pi(k)$ satisfying conditions (9) and (10): if we choose

$$\pi(k) \quad \propto \quad \xi^{k(p-\frac{k+1}{2})} \tag{11}$$

with $C_{\gamma,\tau}^{-1} < \xi < C_{\gamma,M_{\text{bm}}}^{-1}$, it satisfies the conditions. Furthermore, if we choose $\xi = 1$, which leads to

$$\pi(k) \quad = \quad \frac{1}{R_n + 1}, \tag{12}$$

the conditions are met if $\tau > (1+\epsilon_1)(1+\gamma^{-1})^{-1}$ and $\exp(M_{\text{bm}}) > 2(1+\gamma^{-1})^{1/2}$.

Remark. In the sparse linear regression literature, a common choice for the prior on the unknown sparsity k is $\pi(k) \propto p^{-ck}$ for some constant $c > 0$. See Castillo et al. (2015), Yang et al. (2016) and Martin et al. (2017). If we adopt this type of the prior into the bandwidth selection problem, a naive approach is using $\pi(k) \propto p^{-ck}$ for each row of the Cholesky factor: it results in $\pi(k) \propto p^{-ck(p-k)}$. To obtain strong model selection consistency, in this case, M_{bm} in condition (A2) has to be $M_{\text{bm}} = M'_{\text{bm}} \log p$ for some constant $M'_{\text{bm}} > 0$. Thus, it unnecessarily requires stronger beta-min condition, which can be avoided by using $\pi(k)$ like (11) or (12).

3 Main Results

3.1 Bandwidth Selection Consistency

When there is a natural ordering in the data set, estimating the bandwidth of the precision matrix is important for detecting the dependence structure. It is a crucial first step for the subsequent analysis. In this subsection, we show the bandwidth selection consistency of the proposed prior. Theorem 3.1 states that the posterior distribution puts a mass tending to one at the true bandwidth k_0. Thus, we can detect the true bandwidth using the marginal posterior distribution for the bandwidth k. We call this property the bandwidth selection consistency.

Theorem 3.1. *Consider model (1) and priors (4)–(6). If conditions (A1)–(A4) are satisfied, then we have*

$$\mathbb{E}_0\Big[\pi\big(k \neq k_0 \mid \mathbf{X}_n\big)\Big] \quad = \quad o(1).$$

Informed readers might be aware of the recent work of Cao et al. (2019) considering the selection of sparse Cholesky factors. It should be noted that their method is not applicable to the bandwidth selection problem. The key issue is that their method is not adaptive to the unknown sparsity corresponding to the true bandwidth k_0 in this paper: to obtain the selection consistency, the choice of hyperparameter should depend on k_0, which is unknown and of interest. Furthermore, they required stronger conditions in terms of dimensionality p, true sparsity k_0, eigenvalues of the true precision matrix and beta-min for the strong model selection consistency.

Remark. The bandwidth selection result does not necessarily imply the consistency of the Bayes factor. Note that prior (4), $\pi(d_j) \propto d_j^{-1}$ and

$$\pi(k) \quad \propto \quad \prod_{j=2}^{p} \left(\widehat{d}_j^{(k)}\right)^{\frac{\tau n}{2}}, \tag{13}$$

and priors (4), (5) and $\pi(k) \propto 1$ lead to the same marginal posterior for k. Thus, the above priors also achieve the bandwidth selection consistency in Theorem 3.1. However, (13) might be inappropriate when the Bayes factor is of interest, because the ratio of normalizing terms induced by prior (13) (C_0 and C_1 in (14)) have a non-ignorable effect on the Bayes factor.

3.2 Consistency of One-Sample Bandwidth Test

In this subsection, we focus on constructing a Bayesian bandwidth test for the testing problem $H_0 : k \leq k^\star$ versus $H_1 : k > k^\star$ for some given $k^\star$. A Bayesian hypothesis test is based on the Bayes factor $B_{10}(\mathbf{X}_n)$ defined by the ratio of marginal likelihoods,

$$B_{10}(\mathbf{X}_n) \quad = \quad \frac{p(\mathbf{X}_n \mid H_1)}{p(\mathbf{X}_n \mid H_0)}.$$

We are interested in the consistency of the Bayes factor which is one of the most important asymptotic properties of the Bayes factor (Dass and Lee, 2004). A Bayes factor is said to be *consistent* if $B_{10}(\mathbf{X}_n)$ converges to zero in probability under the true null hypothesis H_0 and $B_{10}(\mathbf{X}_n)^{-1}$ converges to zero in probability under the true alternative hypothesis H_1.

Although the Bayes factor plays a crucial role in the Bayesian variable selection, its asymptotic behaviors in the high-dimensional setting are not well-understood (Moreno et al., 2010). Few works studied the consistency of the Bayes factor in the high-dimensional settings (Moreno et al., 2010; Wang and Sun, 2014; Wang et al., 2016), which only focused on the pairwise consistency of the Bayes factor. They considered the testing problem $H_0 : k = k^{(0)}$ versus $H_1 : k = k^{(1)}$ for any $k^{(0)} < k^{(1)}$, where k is the number of nonzero elements of the linear regression coefficient. Note that a Bayes factor is said to be *pairwise consistent* if the Bayes factor $B_{10}(\mathbf{X}_n)$ is consistent for any pair of simple hypotheses H_0 and H_1.

We focus on the composite hypotheses $H_0 : k \leq k^\star$ and $H_1 : k > k^\star$ rather than simple hypotheses. To conduct a Bayesian hypothesis test, prior distributions for

both hypotheses should be determined. Denote the prior under the hypothesis H_i as $\pi_i(A_n, D_n, k)$ for $i = 0, 1$. Since the difference between two hypotheses comes only from the bandwidth, we will use the same conditional priors for A_n and D_n given k, i.e. $\pi_i(A_n, D_n, k) = \pi_i(k)\,\pi(A_n, D_n \mid k)$ for $i = 0, 1$, where $\pi(A_n, D_n \mid k)$ is chosen as (4) and (5). We suggest using priors $\pi_0(k)$ and $\pi_1(k)$ such that

$$\begin{aligned} \pi_0(k) &= C_0^{-1}\pi(k), \quad k = 0, 1, \ldots, k^\star, \\ \pi_1(k) &= C_1^{-1}\pi(k), \quad k = k^\star + 1, \ldots, R_n, \end{aligned} \tag{14}$$

where $C_0 = \sum_{k=0}^{k^\star} \pi(k)$ and $C_1 = \sum_{k=k^\star+1}^{R_n} \pi(k)$. Then, the Bayes factor has the following analytic form,

$$\begin{aligned} B_{10}(\mathbf{X}_n) &= \frac{\sum_{k>k^\star} \int p(\mathbf{X}_n \mid \Omega_n, k)\pi(\Omega_n \mid k)\pi_1(k)d\Omega_n}{\sum_{k\le k^\star} \int p(\mathbf{X}_n \mid \Omega_n, k)\pi(\Omega_n \mid k)\pi_0(k)d\Omega_n} \\ &= \frac{\pi(k > k^\star \mid \mathbf{X}_n)}{\pi(k \le k^\star \mid \mathbf{X}_n)} \times \frac{C_0}{C_1}, \end{aligned}$$

where the marginal posterior $\pi(k \mid \mathbf{X}_n)$ is given in (7) up to some normalizing constant. Note that, the Bayes factor can be defined because both hypotheses have the same improper priors on D_n. We will show that the Bayes factor is consistent for any composite hypotheses $H_0 : k \le k^\star$ and $H_1 : k > k^\star$, which is generally *stronger than the pairwise consistency* of the Bayes factor. If we assume that $\pi_1(k)/\pi_0(k') = O(1)$ for any k and k', then one can see that the consistency of the Bayes factor for hypotheses $H_0 : k \le k^\star$ and $H_1 : k > k^\star$ for any $k^\star$ implies the pairwise consistency of the Bayes factor for any pair of simple hypotheses $H_0 : k = k^{(0)}$ and $H_1 : k = k^{(1)}$ for $k^{(0)} < k^{(1)}$.

For given positive constants M_{bm}, γ and $\tau \in (0, 0.4]$ and integers R_n, k_0 and $k^\star$, define

$$\begin{aligned} T_{n,H_0,k_0,k^\star} &= k^\star \cdot \left\{C_{\gamma,\tau}^{-1}\right\}^{(k^\star+1-k_0)\cdot(p-\frac{R_n+k_0+1}{2})}, \\ T_{n,H_1,k_0,k^\star} &= (R_n - k^\star) \cdot \left\{C_{\gamma,M_{\mathrm{bm}}}\right\}^{(k_0-k^\star)\cdot(p-\frac{k^\star+k_0+1}{2})}, \end{aligned}$$

where $C_{\gamma,\tau}$ and $C_{\gamma,M_{\mathrm{bm}}}$ are defined in condition (A4). Theorem 3.2 shows the convergence rates of Bayes factors under each hypothesis. It turns out that $\pi(k) = 1/(R_n + 1)$ is sufficient for the consistency of the Bayes factor.

Theorem 3.2. *Consider model* (1) *and hypothesis testing problem* $H_0 : k \le k^\star$ *versus* $H_1 : k > k^\star$. *Assume priors* (4) *and* (5) *for* $\pi(A_n, D_n \mid k)$ *and the bandwidth priors in* (14) *with* $\pi(k) = 1/(R_n + 1)$. *If conditions (A1)–(A3) hold,* $\tau > \gamma(1 + \epsilon_1)/(1 + \gamma)$ *and* $\exp(M_{\mathrm{bm}}) > 2\{(1 + \gamma)/\gamma\}^{1/2}$, *then the Bayes factor* $B_{10}(\mathbf{X}_n)$ *is consistent under* $\mathbb{P}_0$. *Moreover, under* $H_0 : k \le k^\star$, *we have*

$$B_{10}(\mathbf{X}_n) = O_p\big(T_{n,H_0,k_0,k^\star}\big),$$

and under $H_1 : k > k^\star$,

$$B_{10}(\mathbf{X}_n)^{-1} = O_p(T_{n,H_1,k_0,k^\star}).$$

Remark. By choosing the prior $\pi(k) = 1/(R_n + 1)$, it implies that $C_{\gamma,\tau}^{-1}$ and $C_{\gamma,M_{\text{bm}}}$ are strictly smaller than 1 by (11). Since we assume that $p \to \infty$ as $n \to \infty$, one can easily check that $T_{n,H_0,k_0,k^\star}$ and $T_{n,H_1,k_0,k^\star}$ go to zero as $n \to \infty$ under $H_0 : k \leq k^\star$ and $H_1 : k > k^\star$, respectively.

Remark. Note that if we use prior (11) with $\xi \neq 1$, the effect of the prior, C_0/C_1, can dominate the posterior ratio, $\pi(k > k^\star \mid \mathbf{X}_n)/\pi(k \leq k^\star \mid \mathbf{X}_n)$ in the Bayes factor. Because the prior knowledge on the bandwidth is usually not sufficient, it is clearly undesirable. Moreover, the direction of effect is the opposite of the prior knowledge.

An et al. (2014) and Cheng et al. (2017) developed frequentist bandwidth tests for the hypotheses $H_0 : k \leq k^\star$ versus $H_1 : k > k^\star$ and showed that their test statistic is asymptotically normal under the null and has a power converging to one as $n \wedge p \to \infty$. Compared with the result in Theorem 3.2, Cheng et al. (2017) required the upper bound $k_0 = O([n/\log p]^{1/2})$ for the true bandwidth k_0, which is much stronger than our condition (A3). An et al. (2014) allowed $k_0 \leq n - 4$, but assumed that the partial correlation coefficient between X_{ij} and $X_{i,j-k_0}$ given $X_{i,j-k_0+1}, \ldots, X_{i,j-1}$ is of order $o(n^{-1})$. It implies that $\max_j |a_{0,jj-k_0}|$ converges to zero at some rate. Thus, the nonzero elements $a_{0,jj-k_0}$, $j = k_0 + 1, \ldots, p$ are required to converge to zero. This is different from our condition (A2), which does not require the nonzero elements including $a_{0,jj-k_0}$, $j = k_0 + 1, \ldots, p$ to converge to zero, also they can.

Johnson and Rossell (2010, 2012) and Rossell and Rubio (2018) pointed out that the use of local alternative prior leads to imbalanced convergence rates for the Bayes factors, and showed that this issue can be avoided by using non-local alternative priors. However, interestingly, convergence rates for the Bayes factors in Theorem 3.2 yield *similar order of rates* under both hypotheses without using a non-local prior. Roughly speaking, the imbalance issue can be ameliorated by introducing the beta-min condition (Condition (A2)). To simplify the situation, consider the model

$$Y \quad = \quad X\beta^{(k)} + \epsilon,$$

where $Y = (Y_1, \ldots, Y_n)^T$, $X \in \mathbb{R}^{n \times p}$, $\beta^{(k)} = (\beta_1, \ldots, \beta_k, 0, \ldots, 0)^T \in \mathbb{R}^p$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Suppose priors (4) and (5) are imposed on $(\beta_1, \ldots, \beta_k)^T$ and σ^2 given k. Consider hypotheses $H_0 : k = k_1$ and $H_1 : k = k_2$, where $k_1 < k_2$, and assume that the eigenvalues of $X_{(1:k_2)}$ are bounded and $k_2 - k_1 \to \infty$ as $n \to \infty$ for simplicity. Note that the prior for $\beta^{(k_2)}$ is a local alternative prior because $\pi(\beta^{(k_2)} \mid \sigma^2) > c$ on $\{\beta^{(k_2)} \in \mathbb{R}^{k_2} : \beta^{(k_2)} = (\beta_1, \ldots, \beta_{k_1}, 0, \ldots, 0)^T\}$ for some constant $c > 0$. If H_0 is true, $B_{10}(Y)$ decreases at rate $O_p(e^{-c_0(k_2-k_1)})$ for some constant $c_0 > 0$ based on techniques in the proof of Theorem 3.1. On the other hand, if H_1 is true, $B_{10}(Y)^{-1}$ decreases exponentially with $n(k_2 - k_1)\beta_{\min}^2$, where $\beta_{\min}$ is the lower bound for the absolute of nonzero elements of $\beta_0^{(k_2)}$. Johnson and Rossell (2010, 2012) and Rossell and Rubio (2018) assumed that $\beta_{\min}^2 > c_1$ for some constant $c_1 > 0$. In that case, $B_{10}(Y)^{-1}$ decreases exponentially with $n(k_2 - k_1)c_1$, which causes the imbalanced convergence rates. However, if we assume $\beta_{\min}^2 \geq c_2 n^{-1}$ similar to condition (A2), $B_{10}(Y)^{-1}$ decreases at rate $O_p(e^{-c_2(k_2-k_1)})$ for some constant $c_2 > 0$. Thus, convergence rates for the Bayes factors have similar order under the both hypotheses.

The above argument does not mean that the non-local priors are not useful for our problem. We note that the balanced convergence rates by using the beta-min condition is different from those by using the non-local prior. The former makes the rate of $B_{10}(Y)^{-1}$ slower under H_1, while the latter makes the rate of $B_{10}(Y)$ faster under H_0. Thus, the use of non-local priors might improve the rates of convergence for $B_{10}(Y)$ under H_0 in Theorem 3.2. However, it will increase the computational burden and is unclear which rate one can achieve using the non-local prior under condition (A2), so we leave it as a future work.

3.3 Consistency of Two-Sample Bandwidth Test

Suppose we have two data sets from the models

$$\begin{aligned} X_1, \ldots, X_{n_1} \mid \Omega_{1n_1} &\overset{i.i.d.}{\sim} N_p(0, \Omega_{1n_1}^{-1}), \\ Y_1, \ldots, Y_{n_2} \mid \Omega_{2n_2} &\overset{i.i.d.}{\sim} N_p(0, \Omega_{2n_2}^{-1}), \end{aligned} \tag{15}$$

where $\Omega_{1n_1} = (I_p - A_{1n_1})^T D_{1n_1}^{-1}(I_p - A_{1n_1})$ and $\Omega_{2n_2} = (I_p - A_{2n_2})^T D_{2n_2}^{-1}(I_p - A_{2n_2})$ are the MCDs. Denote the bandwidth of Ω_{in_i} as k_i for $i = 1, 2$. In this subsection, our interest is the test of equality between two bandwidths k_1 and k_2, the *two-sample bandwidth test*. We consider the hypothesis testing problem $H_0 : k_1 = k_2$ versus $H_1 : k_1 \neq k_2$ and investigate the asymptotic behavior of the Bayes factor,

$$B_{10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}) \quad = \quad \frac{p(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2} \mid H_1)}{p(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2} \mid H_0)},$$

where $\mathbf{X}_{n_1} = (X_1^T, \ldots, X_{n_1}^T)^T \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{Y}_{n_2} = (Y_1^T, \ldots, Y_{n_2}^T)^T \in \mathbb{R}^{n_2 \times p}$. Suppose that multivariate observations are collected from two populations, and a test of the equality of dependence structure is the main interest. When the dependence structure is directly related to how many previous variables influencing the current variable, two-sample bandwidth test provides a suitable answer.

Denote the priors under H_0 and H_1 as, respectively

$$\begin{aligned} & \pi_0\big(A_{1n_1}, D_{1n_1}, A_{2n_2}, D_{2n_2} \mid k\big) \\ = \; & \pi\big(A_{1n_1}, D_{1n_1} \mid k\big)\, \pi\big(A_{2n_2}, D_{2n_2} \mid k\big)\, \pi_0(k), \quad k = 0, 1, \ldots, R_n, \end{aligned}$$

and

$$\begin{aligned} & \pi_1\big(A_{1n_1}, D_{1n_1}, A_{2n_2}, D_{2n_2} \mid k_1, k_2\big) \\ = \; & \pi\big(A_{1n_1}, D_{1n_1} \mid k_1\big)\, \pi\big(A_{2n_2}, D_{2n_2} \mid k_2\big)\, \pi_1(k_1, k_2), \quad 0 \leq k_1 \neq k_2 \leq R_n. \end{aligned}$$

We suggest the following conditional priors $\pi(A_{1n_1}, D_{1n_1} \mid k_1)$ and $\pi(A_{2n_2}, D_{2n_2} \mid k_2)$

for any given k_1 and k_2,

$$\begin{aligned} a_{1,j}^{(k_1)} \mid d_{1,j}, k_1 &\overset{ind}{\sim} N_{k_{1j}}\Big(\widehat{a}_{1,j}^{(k_1)}, \frac{d_{1,j}}{\gamma}\big(\mathbf{X}_{j(k_1)}^T\mathbf{X}_{j(k_1)}\big)^{-1}\Big), \\ \pi(d_{1,j}) &\overset{i.i.d.}{\propto} d_{1,j}^{\tau n_1/2-1}, \\ a_{2,j}^{(k_2)} \mid d_{2,j}, k_2 &\overset{ind}{\sim} N_{k_{2j}}\Big(\widehat{a}_{2,j}^{(k_2)}, \frac{d_{2,j}}{\gamma}\big(\mathbf{Y}_{j(k_2)}^T\mathbf{Y}_{j(k_2)}\big)^{-1}\Big), \\ \pi(d_{2,j}) &\overset{i.i.d.}{\propto} d_{2,j}^{\tau n_2/2-1}, \end{aligned} \tag{16}$$

where $k_{ij} = k_i \wedge (j-1)$, $a_{i,j}^{(k_i)} \in \mathbb{R}^{k_{ij}}$ is the nonzero elements in the jth row of A_{in_i} and $D_{in_i} = diag(d_{i,j})$ for $i = 1, 2$. Similar to the previous notations, we denote $\widehat{a}_{1,j}^{(k_1)} = (\mathbf{X}_{j(k_1)}^T\mathbf{X}_{j(k_1)})^{-1}\mathbf{X}_{j(k_1)}^T\tilde{X}_j$ and $\widehat{a}_{2,j}^{(k_2)} = (\mathbf{Y}_{j(k_2)}^T\mathbf{Y}_{j(k_2)})^{-1}\mathbf{Y}_{j(k_2)}^T\tilde{Y}_j$, where $\mathbf{Y}_{j(k_2)} \in \mathbb{R}^{n\times k_2}$ is the sub-matrix consisting of $(j-k_2)_1, \ldots, (j-1)$th columns of $\mathbf{Y}_n$. The priors on bandwidths are chosen as

$$\begin{aligned} \pi_0(k) &= \frac{1}{R_n+1}, \quad k = 0, 1, \ldots, R_n, \\ \pi_1(k_1, k_2) &= \frac{1}{R_n(R_n+1)}, \quad 0 \le k_1 \ne k_2 \le R_n. \end{aligned} \tag{17}$$

This choice of priors leads to an analytic form of the Bayes factor,

$$B_{10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}) = \frac{\sum_{k_1\ne k_2}\pi(k_1 \mid \mathbf{X}_{n_1})\pi(k_2 \mid \mathbf{Y}_{n_2})}{\sum_{k_1=k_2}\pi(k_1 \mid \mathbf{X}_{n_1})\pi(k_2 \mid \mathbf{Y}_{n_2})} \times R_n^{-1},$$

where the marginal posterior distributions $\pi(k_1 \mid \mathbf{X}_{n_1})$ and $\pi(k_2 \mid \mathbf{Y}_{n_2})$ are known up to some normalizing constants similar to (7). We denote Ω_{0,in_i} as the true precision matrix with bandwidth k_{0i} for $i = 1, 2$ and assume that p tends to infinity as $n = n_1 \wedge n_2 \to \infty$. Theorem 3.3 gives a sufficient condition for the consistency of the Bayes factor $B_{10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2})$ by calculating the convergence rates.

Theorem 3.3. *Consider model* (15) *and hypotheses* $H_0 : k_1 = k_2$ *and* $H_1 : k_1 \ne k_2$. *Assume the conditional priors given bandwidths* (16) *and the bandwidth priors* (17). *If conditions (A1)–(A3) for* $\Omega_{0,1n_1}$, $\Omega_{0,2n_2}$ *and priors are satisfied,* $\tau > \gamma(1+\epsilon_1)/(1+\gamma)$ *and* $\exp(M_{\text{bm}}) > 2\{(1+\gamma)/\gamma\}^{1/2}$, *then the Bayes factor* $B_{10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2})$ *is consistent under* $\mathbb{P}_0$. *Moreover, under* $H_0 : k_1 = k_2$, *we have*

$$B_{10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}) = O_p\left(\frac{k_0}{R_n - k_0}T_{n,H_1,k_0,k_0-1} + \frac{R_n - k_0}{k_0}T_{n,H_0,k_0,k_0}\right),$$

and under $H_1 : k_1 \ne k_2$,

$$B_{10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2})^{-1} = O_p\left(\frac{R_n\, k_{\min}}{R_n - k_{\min}}T_{n,H_1,k_{\min},k_{\min}-1} + \frac{R_n(R_n - k_{\min})}{k_{\min}}T_{n,H_0,k_{\min},k_{\min}}\right),$$

where $k_{\min} = k_{01} \wedge k_{02}$.

As mentioned earlier, to the best of our knowledge, this is the first consistent two-sample bandwidth test result in high-dimensional settings. Frequentist testing procedures in An et al. (2014) and Cheng et al. (2017) focused only on the one-sample bandwidth test, and it is unclear whether these methods can be extended to the two-sample testing problem.

Note that the hypothesis testing problem $H_0 : k_1 = k_2$ versus $H_1 : k_1 \neq k_2$ is different from the hypothesis testing $H_0 : \Omega_{1n_1} = \Omega_{2n_2}$ versus $H_1 : \Omega_{1n_1} \neq \Omega_{2n_2}$ in Cai et al. (2013). The latter testing problem is called the two-sample precision (or covariance) test. The two-sample bandwidth test is weaker than the two-sample precision test, i.e. if the two-sample bandwidth test supports the null hypothesis, then one can further conduct the two-sample precision test.

4 Numerical Results

We have proved the bandwidth selection consistency and convergence rates of Bayes factors based on priors (4)–(6). In this section, we conduct simulation studies to describe the practical performance of the proposed method. Throughout the section, we use the prior $\pi(k) = 1/(R_n + 1)$.

4.1 Comparison with Other Bandwidth Tests

In this subsection, we compared the performance of our method with those of other bandwidth selection procedures. Since we have bandwidth selection consistency (Theorem 3.1), we suggest using the posterior mode to estimate the true bandwidth k_0. We chose the bandwidth test of An et al. (2014) as a frequentist competitor and the bandwidth selection procedures of Banerjee and Ghosal (2014) and Lee and Lee (2018a) as Bayesian competitors. Significance levels for bandwidth tests in An et al. (2014) were varied $\alpha = 0.001, 0.005, 0.01$, but only the result with $\alpha = 0.01$ are reported since they gave similar results. For Banerjee and Ghosal (2014) and Lee and Lee (2018a), we used the prior $\pi(k) \propto \exp(-k^4)$ as they suggested. Note that these Bayesian procedures do not guarantee the bandwidth selection consistency.

To calculate the marginal posterior in (7), the hyperparameters γ, τ and R_n should be determined. As a pragmatic approach, we incorporated cross-validation (CV) to select γ, and fixed $\tau = 0.01$ and $R_n = \lfloor n/4 \rfloor$; in our experiments, we also tried $\tau = 0.01, 0.02, \ldots, 0.30$ and selected τ via CV, but found that $\tau = 0.01$ is selected in most cases. Since condition (A3) requires $R_n \leq n\,\tau\epsilon_1(1+\epsilon_1)^{-1}$ for some small ϵ_1 and $\tau > 0$ for asymptotic results, the choice $R_n = \lfloor n/4 \rfloor$ seems reasonable in finite samples. We randomly divided the data $\mathbf{X}_n$ into two subsamples, the test set $\mathbf{X}^{te}_{n_1}$ and training set $\mathbf{X}^{tr}_{n_2}$, where $n_1 = \lceil n/3 \rceil$ and $n_2 = n - n_1$. Let $\widehat{k}(\gamma)$ be the posterior mode based on $\mathbf{X}^{tr}_{n_2}$ and a given γ, and define the mean squared error (MSE)

$$MSE(\gamma) \quad = \quad \sum_{j=1}^{p} \|\tilde{X}^{te}_j - \mathbf{X}^{te}_{j(\widehat{k}(\gamma))} \widehat{a}_j^{(\widehat{k}(\gamma))}\|_2^2.$$

	$(\widehat{p}_0,\ \widehat{k}_0)$			
	$(n=70, p=100)$	$(n=70, p=200)$	$(n=200, p=100)$	$(n=200, p=200)$
BBS	(0.96, 4.94)	(0.98, 4.98)	(1.00, 5.00)	(1.00, 5.00)
BA1	(0.78, 4.84)	(0.96, 5.16)	(0.96, 5.06)	(1.00, 5.00)
BA2	(0.78, 4.84)	(0.98, 5.18)	(0.96, 5.06)	(1.00, 5.00)
LL	(0.00, 1.00)	(0.00, 1.00)	(0.00, 1.02)	(0.00, 1.04)
BG	(0.00, 1.00)	(0.00, 1.00)	(0.00, 1.00)	(0.00, 1.00)

Table 1: The summary statistics for each setting are represented, where $k_0 = 5$ and $[A_{0,\min}, A_{0,\max}] = [0.1, 0.1]$. BBS: the proposed method in this paper. LL: bandwidth selection procedure of Lee and Lee (2018a). BG: bandwidth selection procedure of Banerjee and Ghosal (2014). BA1 and BA2: algorithms 1 and 2 in An et al. (2014), respectively.

	$(\widehat{p}_0,\ \widehat{k}_0)$			
	$(n=70, p=100)$	$(n=70, p=200)$	$(n=200, p=100)$	$(n=200, p=200)$
BBS	(0.94, 10.02)	(0.96, 10.06)	(1.00, 10.00)	(1.00, 10.00)
BA1	(0.72, 9.66)	(0.98, 10.14)	(1.00, 10.00)	(0.98, 10.06)
BA2	(0.74, 9.72)	(0.98, 10.14)	(1.00, 10.00)	(0.96, 10.08)
LL	(0.00, 2.12)	(0.00, 2.86)	(0.00, 3.36)	(0.00, 4.00)
BG	(0.00, 1.00)	(0.00, 1.00)	(0.00, 1.00)	(0.00, 1.00)

Table 2: The summary statistics for each setting are represented, where $k_0 = 10$ and $[A_{0,\min}, A_{0,\max}] = [0.1, 0.2]$.

We divided the data 20 times and selected $\widehat{\gamma}$ which minimizes $20^{-1}\sum_{\nu=1}^{20} MSE_\nu(\gamma)$ among $0.01, 0.02, \ldots, 0.40$, where $MSE_\nu(\gamma)$ is the MSE from the νth subsampling. Depending on the purpose of the analysis, other criteria besides MSE can be adopted.

The data sets were generated from $N_p(0, \Omega_{0n}^{-1})$, where $\Omega_{0n} = (I_p - A_{0n})^T D_{0n}^{-1}(I_p - A_{0n})$. For each $j = k_0 + 1, \ldots, p$, nonzero elements in the jth row of the true Cholesky factor A_{0n} were sampled from $Unif(A_{0,\min}, A_{0,\max})$ and ordered to satisfy $a_{0,jl} \leq a_{0,jl'}$ for any $l < l'$. The diagonal elements of D_{0n} were generated from $Unif(5, 10)$. To investigate performance in various settings, the values of n, p, k_0, $A_{0,\min}$ and $A_{0,\max}$ were varied. The simulation results, based on 50 simulated data sets for each setting, are reported in Table 1 and Figure 1. We denoted the proposed method in this paper as BBS, the Bayesian Bandwidth Selector.

Performance of each method were evaluated by the proportion of correct detections of k_0, $\widehat{p}_0 = \sum_{s=1}^{50} I(\hat{k}_0^{(s)} = k_0)/50$, and averaged bandwidth estimate, $\widehat{k}_0 = \sum_{s=1}^{50} \hat{k}_0^{(s)}/50$, where $\hat{k}_0^{(s)}$ is the estimated bandwidth for the sth data set. Our method, the BBS, consistently outperformed other competitors in most settings. The bandwidth selection procedures of An et al. (2014) worked reasonably well for large n and large p cases, but it seems somewhat unstable when $(n = 70, p = 100)$. Although Lee and Lee (2018a) is slightly better than Banerjee and Ghosal (2014), both of them consistently underestimated the true bandwidth k_0. The proposed prior $\pi(k) \propto \exp(-k^4)$ seems to be too strong to put sufficient masses near the true bandwidth k_0 especially when k_0 is

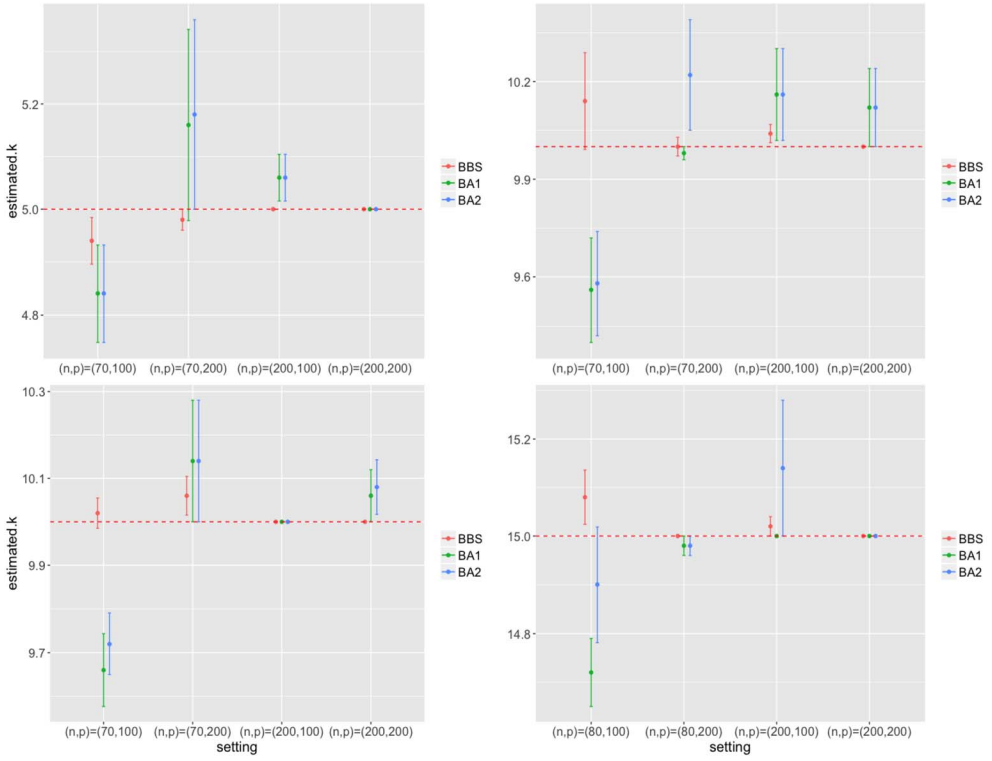

Figure 1: The summary plots for estimated bandwidth in various settings. The middle dot and bar represent the mean and standard error of the mean, respectively, based on 50 simulations. $[A_{0,\min}, A_{0,\max}] = [0.1, 0.1]$ was used for the top row, while $[A_{0,\min}, A_{0,\max}] = [0.1, 0.2]$ was used for the bottom row. The red dashed line is the true bandwidth.

not small. Figure 1 shows the bandwidth selection results of BBS and the test in An et al. (2014) to compare the performance of the two methods at a glance. As shown in Tables 1 and 2, the BBS outperformed the bandwidth tests in An et al. (2014) in most cases.

4.2 Simulation Study for Two-Sample Bandwidth Test

In this section, we demonstrate the performance of the proposed two-sample bandwidth test in various simulation cases. We generated $X_1, \ldots, X_{n_1} \overset{i.i.d.}{\sim} N_p(0, \Omega_{0,1n_1}^{-1})$ and $Y_1, \ldots, Y_{n_2} \overset{i.i.d.}{\sim} N_p(0, \Omega_{0,2n_2}^{-1})$, where $\Omega_{0,1n_1} = (I_p - A_{0,1n_1})^T D_{0,1n_1}^{-1}(I_p - A_{0,1n_1})$ and $\Omega_{0,2n_2} = (I_p - A_{0,2n_2})^T D_{0,2n_2}^{-1}(I_p - A_{0,2n_2})$. Nonzero elements in each row of two true Cholesky factors were generated from $Unif(A_{01,\min}, A_{01,\max})$ and $Unif(A_{02,\min}, A_{02,\max})$, respectively. The diagonal elements of $D_{0,1n_1}$ and $D_{0,2n_2}$ were generated from

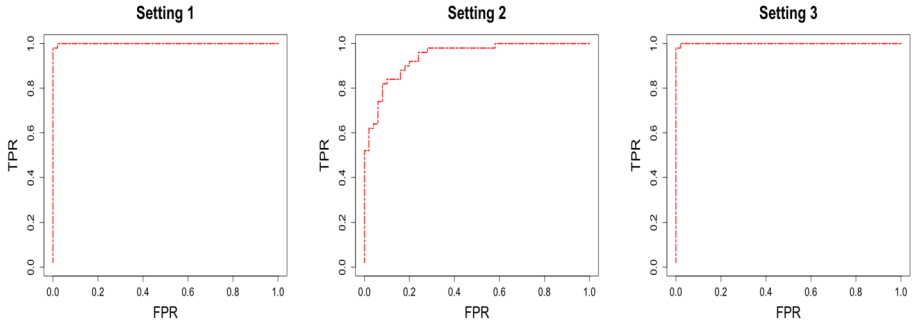

Figure 2: ROC curves for the proposed two-sample test in this paper under various settings.

$Unif(2,5)$. Throughout the simulations, the hyperparameters τ and R_n were set at $\tau = 0.01$ and $R_n = \lfloor \min(n_1, n_2)/4 \rfloor$, and the values of γ for two groups were chosen by CV method described in Section 4.1.

As the first setting, we chose $n_1 = n_2 = 50$, $p = 100$, $A_{01,\min} = 0.1$, $A_{01,\max} = 0.2$, $A_{02,\min} = 0.2$ and $A_{02,\max} = 0.4$. The true bandwidths k_{01} and k_{02} were set at $k_{01} = k_{02} = 2$ under the true null setting, and set at $k_{01} = 2$ and $k_{02} = 3$ under the true alternative setting. Simulated data $\mathbf{X}_{n_1}$ and $\mathbf{Y}_{n_2}$ were generated 50 times under both true null and alternative settings. To illustrate the performance of the two-sample bandwidth test, receiver operating characteristic (ROC) curve is given at Figure 2 with the label "Setting 1". The points of the curve were obtained based on various choices of thresholds for the log Bayes factor. The ROC curve shows almost perfect performance. We also tried $n_1 = 50$, $n_2 = 100$, $p = 150$, $A_{01,\min} = 0.1$, $A_{01,\max} = 0.2$, $A_{02,\min} = 0.2$ and $A_{02,\max} = 0.4$. The true bandwidths k_{01} and k_{02} were set at $k_{01} = k_{02} = 10$ under the true null setting, and set at $k_{01} = 10$ and $k_{02} = 11$ under the true alternative setting. The ROC curve based on 50 simulated data from each setting is given at Figure 2 with the label "Setting 2". The performance is slightly worse than the first setting because of the large dimensionality and relatively small n_1, but still seems reasonable. In fact, as the third setting, we tried the same setting with "Setting 2" except $n_1 = 100$. The ROC curve is given at Figure 2 with the label "Setting 3", which shows almost perfect performance.

4.3 Telephone Call Center Data

We illustrate the performance of the proposed method using the telephone call center data previously analyzed by Huang et al. (2006), Bickel and Levina (2008) and An et al. (2014). The phone calls were recorded from 7:00 am until midnight from a call center of a major U.S. financial organization. The data were collected for 239 days in 2002 except holidays, weekends and days when the recording system did not work properly. The number of calls were counted for every 10 minutes, and a total of 102 intervals were obtained on each day. We denote the number of calls on the jth time interval of the

ith day as N_{ij} for each $i = 1, \ldots, 239$ and $j = 1, \ldots, 102$. As in Huang et al. (2006), Bickel and Levina (2008) and An et al. (2014), a transformation $X_{ij} = \sqrt{N_{ij} + 1/4}$ was applied to make the data close to the random sample from normal distribution. The transformed data set was centered. For more details about the data set, see Huang et al. (2006).

We are interested in predicting the number of phone calls from the 52nd to 102nd time intervals using the previous counts on each day. The best linear predictor of X_{ij} from $X_i^j = (X_{i1}, \ldots, X_{i,j-1})^T$,

$$\widehat{X}_{ij} \quad = \quad \mu_j + \Sigma_{(j,1:(j-1))} \left[\Sigma_{(1:(j-1),1:(j-1))}\right]^{-1} (X_i^j - \mu^j), \tag{18}$$

was used to predict X_{ij} for each $j = 52, \ldots, 102$, where $\mu_j = \mathbb{E}(X_{1j})$, $\mu^j = (\mu_1, \ldots, \mu_{j-1})^T$ and Σ_{S_1,S_2} is a sub-matrix of Σ consisting of the S_1th rows and the S_2th columns for given index sets S_1 and S_2. We used the first 205 days ($i = 1, \ldots, 205$) as a training set and the last 34 days ($i = 206, \ldots, 239$) as a test set. To calculate the best linear predictor (18), the unknown parameters are need to be estimated. Because it is reasonable to assume the existence of the natural (time) ordering, we plugged the estimators $\widehat{\mu}^j = \sum_{i=1}^{205} X_i^j/205$ and $\widehat{\Sigma}_k = \left\{(I_p - \widehat{A}_{nk})^T \widehat{D}_{nk}^{-1} (I_p - \widehat{A}_{nk})\right\}^{-1}$ into (18), where $\widehat{A}_{nk}$ and $\widehat{D}_{nk}$ are estimators based on the training set.

We applied the proposed methods in this paper, An et al. (2014) and Bickel and Levina (2008) to estimate the bandwidth k using the training set, and compared the prediction errors $\mathrm{PE}_j = \sum_{i=206}^{239} |\widehat{X}_{ij} - X_{ij}|/34$ for each $j = 52, \ldots, 102$. We defined the average of prediction errors, $\sum_{j=52}^{102} \mathrm{PE}_j/51$ to illustrate the performance of estimated bandwidths. For a fair comparison, we used the same estimator $\widehat{\Sigma}_k$ and only chose different bandwidths depending on the selection procedure. Since the goal of the analysis is prediction, the average of prediction errors using training set was used as the criterion for CV. The hyperparameters γ and τ minimizing the average of prediction errors were chosen among $\gamma = 0.01, 0.02, \ldots, 0.40$ and $\tau = 0.01, 0.02, \ldots, 0.30$. We used $R_n = \lfloor n/4 \rfloor$ where $n = 205$ since the number of training set is 205, but the larger values of R_n gave the same result. Based on the selected hyperparameters, our method, the BBS, gave the estimated bandwidth $\widehat{k} = 5$. Algorithms 1 and 2 with $\alpha = 0.01$ in An et al. (2014) determined the bandwidth as 8 and 10, respectively, and Bickel and Levina (2008) selected the bandwidth as 19 based on a resampling scheme proposed in their paper. The average of prediction errors were 0.5347, 0.5474, 0.5568 and 0.5609 at bandwidth $k = 5$, 8, 10 and 19, respectively. Note that if we use the sample covariance matrix instead of the banded estimator $\widehat{\Sigma}_k$, it gives the average prediction error 0.7008. Thus, the banded estimator of Σ benefits in this case, and our bandwidth estimate yields smaller average prediction error compared with other procedures. Figure 3 represents the averages of prediction errors for various bandwidth values k. The minimum error is attained at $k = 4$. None of the above methods achieves the optimal bandwidth $k = 4$, but the bandwidth obtained from our method is closest to 4.

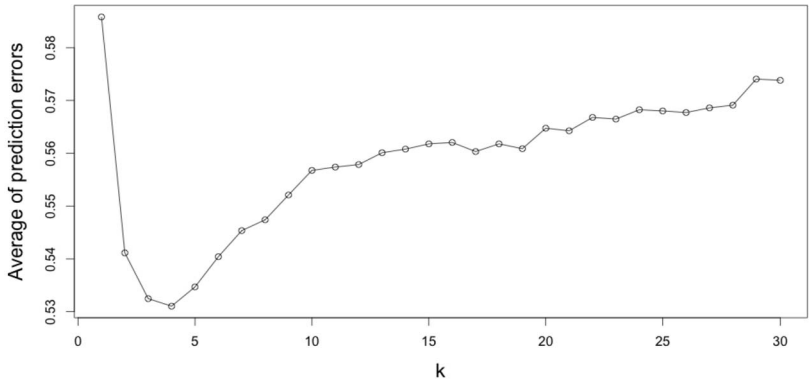

Figure 3: The averages of prediction errors are represented for various bandwidth values k.

5 Discussion

Throughout the paper, we assumed that each row of the Cholesky factor has the same bandwidth for simplicity. It can be extended to more general setting allowing different bandwidth for each row. If we denote the bandwidth for the jth row as $k_0^{(j)}$ and $k_{0,\max} = \max_{1\leq j\leq p} k_0^{(j)}$, then one can conduct the bandwidth test for $k_{0,\max}$. Theoretical results in this paper also hold for the maximum bandwidth $k_{0,\max}$ selection problem with possibly some additional conditions. For example, if $k_0^{(j)} = k_{0,\max}$ except only finite j's, then the proposed priors still achieve the theoretical properties in Section 3.

The bandwidth selection problem for the class of *bandable matrices* considered in Banerjee and Ghosal (2014) can serve as one of the interesting future research directions. Note that it has very different characteristics from these for banded matrices since it may not have the true bandwidth. In the bandable case, bandwidth selection aims to find the optimal bandwidth by minimizing the estimation error with respect to some loss function, and it is well known that the optimal bandwidth depends on the choice of loss function (Cai et al., 2010). Thus, if bandwidth selection of a bandable matrix is of primary interest, the prior distribution should be chosen carefully depending on the loss function.

Supplementary Material

Supplementary to "Bayesian Bandwidth Test and Selection for High-dimensional Banded Precision Matrices" (DOI: 10.1214/19-BA1167SUPP; .pdf).

References

An, B., Guo, J., and Liu, Y. (2014). "Hypothesis testing for band size detection of high-dimensional banded precision matrices." *Biometrika*, 101(2): 477–483. MR3215361. doi: https://doi.org/10.1093/biomet/asu006. 738, 739, 743, 746, 749, 750, 751, 752, 753

Banerjee, S. and Ghosal, S. (2014). "Posterior convergence rates for estimating large precision matrices using graphical models." *Electronic Journal of Statistics*, 8(2): 2111–2137. MR3273620. doi: https://doi.org/10.1214/14-EJS945. 737, 738, 742, 743, 749, 750, 754

Banerjee, S. and Ghosal, S. (2015). "Bayesian structure learning in graphical models." *Journal of Multivariate Analysis*, 136: 147–162. MR3321485. doi: https://doi.org/10.1016/j.jmva.2015.01.015. 737, 742

Bickel, P. J. and Levina, E. (2008). "Regularized estimation of large covariance matrices." *The Annals of Statistics*, 36(1): 199–227. MR2387969. doi: https://doi.org/10.1214/009053607000000758. 737, 752, 753

Cai, T. T., Liu, W., and Xia, Y. (2013). "Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings." *Journal of the American Statistical Association*, 108(501): 265–277. MR3174618. doi: https://doi.org/10.1080/01621459.2012.758041. 749

Cai, T. T., Ma, Z., and Wu, Y. (2015). "Optimal estimation and rank detection for sparse spiked covariance matrices." *Probability Theory and Related Fields*, 161(3–4): 781–815. MR3334281. doi: https://doi.org/10.1007/s00440-014-0562-z. 737

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). "Optimal rates of convergence for covariance matrix estimation." *The Annals of Statistics*, 38(4): 2118–2144. MR2676885. doi: https://doi.org/10.1214/09-AOS752. 737, 754

Cai, T. T. and Zhou, H. H. (2012a). "Minimax estimation of large covariance matrices under ℓ_1-norm." *Statistica Sinica*, 1319–1349. MR3027084. 737

Cao, X., Khare, K., and Ghosh, M. (2019). "Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models." *The Annals of Statistics*, 47(1): 319–348. MR3909935. doi: https://doi.org/10.1214/18-AOS1689. 737, 738, 742, 744

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 742, 743

Cheng, G., Zhang, Z., and Zhang, B. (2017). "Test for bandedness of high-dimensional precision matrices." *Journal of Nonparametric Statistics*, 29(4): 884–902. MR3740724. doi: https://doi.org/10.1080/10485252.2017.1375112. 738, 739, 743, 746, 749

Dass, S. C. and Lee, J. (2004). "A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives." *Journal of Statistical Plan-*

ning and Inference, 119(1): 143–152. MR2018454. doi: https://doi.org/10.1016/S0378-3758(02)00413-5. 744

Fan, J., Fan, Y., and Lv, J. (2008). "High dimensional covariance matrix estimation using a factor model." *Journal of Econometrics*, 147(1): 186–197. MR2472991. doi: https://doi.org/10.1016/j.jeconom.2008.09.017. 737

Gao, C. and Zhou, H. H. (2015). "Rate-optimal posterior contraction for sparse PCA." *The Annals of Statistics*, 43(2): 785–818. MR3325710. doi: https://doi.org/10.1214/14-AOS1268. 737

Hu, A. and Negahban, S. (2017). "Minimax Estimation of Bandable Precision Matrices." In *Advances in Neural Information Processing Systems*, 4895–4903. 737

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). "Covariance matrix selection and estimation via penalised normal likelihood." *Biometrika*, 93(1): 85–98. MR2277742. doi: https://doi.org/10.1093/biomet/93.1.85. 752, 753

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2): 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 746

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107(498): 649–660. MR2980074. doi: https://doi.org/10.1080/01621459.2012.682536. 746

Johnstone, I. M. and Lu, A. Y. (2009). "On consistency and sparsity for principal components analysis in high dimensions." *Journal of the American Statistical Association*, 104(486): 682–693. MR2751448. doi: https://doi.org/10.1198/jasa.2009.0121. 737

Lee, K. and Lee, J. (2018a). "Estimating Large Precision Matrices via Modified Cholesky Decomposition." *Accepted to Statistica Sinica*. doi: https://doi.org/10.5705/ss.202018.0476.. 737, 738, 749, 750

Lee, K. and Lee, J. (2018b). "Optimal Bayesian Minimax Rates for Unconstrained Large Covariance Matrices." *Bayesian Analysis*, 13(4): 1211–1229. MR3855369. doi: https://doi.org/10.1214/18-BA1094. 737

Lee, K. and Lin L. (2019). "Supplementary to "Bayesian Bandwidth Test and Selection for High-dimensional Banded Precision Matrices"." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1167SUPP.. 739

Lee, K., Lee, J., and Lin, L. (2018). "Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models." *Accepted to The Annals of Statistics*. 738, 742

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 741

Martin, R., Mess, R., and Walker, S. G. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. MR3624879. doi: https://doi.org/10.3150/15-BEJ797. 740, 741, 742, 743

Moreno, E., Girón, F. J., and Casella, G. (2010). "Consistency of objective Bayes factors as the model dimension grows." *The Annals of Statistics*, 1937–1952. MR2676879. doi: https://doi.org/10.1214/09-AOS754. 744

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. (2014). "Posterior contraction in sparse Bayesian factor models for massive covariance matrices." *The Annals of Statistics*, 42(3): 1102–1130. MR3210997. doi: https://doi.org/10.1214/14-AOS1215. 737

Ren, Z., Sun, T., Zhang, C.-H., Zhou, H. H., et al. (2015). "Asymptotic normality and optimalities in estimation of large Gaussian graphical models." *The Annals of Statistics*, 43(3): 991–1026. MR3346695. doi: https://doi.org/10.1214/14-AOS1286. 742

Rossell, D. and Rubio, F. J. (2018). "Tractable Bayesian variable selection: beyond normality." *Accepted to Journal of the American Statistical Association*. MR3902243. doi: https://doi.org/10.1080/01621459.2017.1371025. 746

Roverato, A. (2000). "Cholesky decomposition of a hyper inverse Wishart matrix." *Biometrika*, 87(1): 99–112. MR1766831. doi: https://doi.org/10.1093/biomet/87.1.99. 738

Shang, Z. and Clayton, M. K. (2011). "Consistency of Bayesian linear model selection with a growing number of parameters." *Journal of Statistical Planning and Inference*, 141(11): 3463–3474. MR2817355. doi: https://doi.org/10.1016/j.jspi.2011.05.002. 741

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018). "Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings." *Statistica Sinica*, 28(2): 1053. MR3791100. 741

Wang, M., Maruyama, Y., et al. (2016). "Consistency of Bayes factor for nonnested model selection when the model dimension grows." *Bernoulli*, 22(4): 2080–2100. MR3498023. doi: https://doi.org/10.3150/15-BEJ720. 744

Wang, M. and Sun, X. (2014). "Bayes factor consistency for nested linear models with a growing number of parameters." *Journal of Statistical Planning and Inference*, 147: 95–105. MR3151848. doi: https://doi.org/10.1016/j.jspi.2013.11.001. 744

Xiang, R., Khare, K., and Ghosh, M. (2015). "High dimensional posterior convergence rates for decomposable graphical models." *Electronic Journal of Statistics*, 9(2): 2828–2854. MR3439186. doi: https://doi.org/10.1214/15-EJS1084. 737, 742

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). "On the computational complexity of high-dimensional Bayesian variable selection." *The Annals of Statistics*, 44(6): 2497–2532. MR3576552. doi: https://doi.org/10.1214/15-AOS1417. 741, 742, 743

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6: 233–243. MR0881437. 740

Acknowledgments

We are grateful to the reviewers and the Editor for their valuable comments. We thank Baiguo An for providing us the telephone call center data. We gratefully acknowledge the funding support from NSF grants IIS 1663870, DMS CAREER 1654579 and a DARPA grant N66001-17-1-4041. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1F1A1059483).